Journal of Modern Applied Statistical Methods

Volume 4 | Issue 2 Article 27

11-1-2005

Sample Size Selection for Pair-Wise Comparisons Using Information Criteria

Xuemei Pan
University of Maryland, xpan1@umd.edu

C. Mitchell Dayton
University of Maryland, cdayton@umd.edu

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 601-608

1538 – 9472/05/$95.00

Early Scholars

Sample Size Selection for Pair-Wise Comparisons Using Information Criteria

Xuemei Pan C. Mitchell Dayton
University of Maryland

This article provides results for rates of correct identifications of paired-comparison information criteria and Tukey HSD as functions of the pattern of mean differences and of sample size. Therefore, the tables provided are useful for selecting sample sizes in real world applications.

Key words: PCIC, sample size, power, information criteria

Introduction

Model-comparison procedures using information-theoretic criteria such as AIC or BIC provide the basis for attractive alternatives to traditional pairwise comparison procedures such as Tukey HSD tests and its many variations. Known as paired-comparisons information criterion, or PCIC, these methods avoid many of the problems associated with conducting a series of correlated significance tests.

In presenting the theoretical background for PCIC, Dayton (1998) reported a small-scale simulation study that provided some evidence concerning the probability of detecting exactly all true pairwise differences among means from several samples. This is referred to as all-pairs power Ramsey (1978) or as the true-model rate by Cribbie and Keselman (2003). Dayton (1998) found that the all-pairs power for PCIC was found to be generally better than that of HSD. In a much more extensive study of PCIC compared with three step-wise multiple comparison procedures (MCPs), Cribbie and Keselman (2003) reported that "when all population means were not equal… {PCIC}… had significantly higher true-model rates than any of the stepwise MCPs." Similarly, Cribbie (2003) reported a simulation study that compared several conventional multiple comparison procedures with PCIC and concluded that PCIC "…had consistently larger true models rates than did familywise error controlling MCPs."

Information is provided in this article concerning the performance of PCIC with respect to rates of correct identifications of patterns of mean differences as a function of sample size and thus, the results are useful for selecting sample sizes for real world applications. These results supplement the very limited simulation results for minimum sample size requirements for selected power levels provided by Dayton (2003).

Summary of PCIC

For K independent groups, many popular pairwise-comparison procedures compute test statistics for each of the $K(K-1)/2$ unique pairs of means and refer these statistics to an appropriate null distribution. Tukey HSD tests, for example, are based on the studentized range statistic for a span of K means. Thus, $K(K-1)/2$ hypotheses of the form $\mu_k = \mu_{k'}$ for $k \neq k'$ are tested. Among the problems with procedures such as this as cited by Dayton (1998) are:

(1) Some arbitrary technique is necessary to control the family-wise type I error rate for the set of correlated pairwise tests;

Xuemei Pan is a Ph D candidate. Her research interests include latent class modeling and model comparison procedures. E-mail her at xpan1@umd.edu. C. Mitchell Dayton is Professor and Chair. His research interests include experimental design and latent class modeling. Email him at cdayton@umd.edu.

(2) The issues of homogeneity of variance and differential sample size pose problems for many paired-comparison procedures;
(3) Intransitive decisions (e.g., outcomes suggesting mean 1 = mean 2, mean 2 = mean 3, but mean 1 < mean 3) are the rule rather than the exception with typical paired comparison procedures since they entail a series of discrete, pairwise significance tests;
(4) There exists a large variety of competing procedures that differ in how type I error is controlled and consequently, in power.

Dayton (1998) proposed using information-theoretic model-selection criteria such as AIC (Akaike, 1973) or BIC (Schwarz, 1978) for selecting the most appropriate ordering of subsets of means for purposes of interpretation. By considering patterns of mean differences, rather than pair-wise differences, the PCIC approach avoids many of the objections raised above. Furthermore, the interpretation of results is facilitated by PCIC to a much greater degree than by conventional pair-wise comparison procedures.

For K independent means, there are a total of 2^{K-1} patterns of ordered subsets with equal means within subsets. For example, with three groups for which the means are ranked and labeled 1, 2, 3, the $2^2 = 4$ distinct ordered subsets are {123}, {1,23}, {12,3} and {1,2,3}, where a comma is used to separate subsets that are unequal in mean value. The basic approach in PCIC is to compute AIC (or, BIC) for each ordered subset based on appropriate model assumptions. Then, the preferred model for purposes of interpretation is the one that satisfies a min(AIC), or min(BIC), criterion.

Assuming a given model and distributional form for the data (e.g., normal), AIC is computed as $-2\text{Log}_e(L) + 2p$, where p is the number of independent parameters estimated in calculating the likelihood, L, for the observed data. Typically, the additive term, 2p, is viewed as a penalty that reflects the complexity of the model. Similarly, BIC is computed as $-2\text{Log}_e(L) + \text{Log}_e(N)p$ where N is the total sample size. For a model with T subsets of means, p equals T+1 assuming homogeneity of variance for the K groups (see Dayton, 1998; 2003, for discussion of related models without the assumption of homogeneity). For example, for the pattern {1, 2, 34} there are three ordered subsets of means so the value of T is 4. The four parameters that are estimated are the mean of group 1, the mean of group 2, the combined mean of groups 3 and 4 and the pooled variance across the four groups. It should be noted that in computing the likelihood for the data, maximum-likelihood estimates for variances are biased (e.g., use N in the denominator for computing the pooled variance).

AIC does not directly involve the sample size in its computation and, as noted by Bozdogan (1987), lacks certain properties of asymptotic consistency usually associated with increasing sample sizes. Also, since $\text{Log}_e(N)$ is larger than the penalty coefficient, 2 for AIC when N is greater than seven, AIC and BIC may, and often do, result in different orderings of subsets of means with, predictably, simpler models being favored by BIC, although AIC tends to select more complex models (i.e., models with a greater number of subsets of means).

Methodology

The main focus of this research was to provide some guidance for selecting sample sizes for comparisons based on information criteria. Power is not only a function of effect size and sample size but also varies in terms of the population pattern of mean differences. In addition for AIC, but not BIC or other asymptotically consistent methods, there are theoretical maximum power levels with respect to certain patterns of mean differences.

In theory, probabilities for selecting models with larger numbers of subsets of means than the true model can be calculated for AIC using results provided by Bozdogan (1987). These calculations provide the upper limits on power that AIC can attain regardless of sample size (as noted above, AIC is not asymptotically consistent). Therefore, when using AIC it is theoretically possible to choose an over-parameterized model even as the sample size

approaches infinity. Model selection criteria which have this property are sometimes called dimension inconsistent. For example with 5 groups, the maximum powers, in theory, for true models with 1 to 5 clusters of means are: .504, .596, .707, .843, and 1.000, respectively. Thus, for one or two clusters of means there is no sample size that will yield all-pairs power of 2/3 for AIC with 5 groups.

For determining minimum sample size requirements, four sets of conditions were considered:

(1) Number of independent groups: k=3, 4, 5 and 6.

(2) Effect size, f, using Cohen's (1969) definition with small (.1), medium (.25) and large (.4) levels for the corresponding one-way ANOVA design with equally-spaced population means.

(3) Power: .50, .67 and .80 representing low, medium and large values.

(4) Patterns of population means: A variety of patterns were examined as shown in the sample size tables below.

Programming in the matrix language, Gauss (Aptech Systems, Inc., 2002), was used to determine minimum sample size requirements for AIC, BIC and HSD. Data were generated by using 1,000 pseudo-random, homoscedastic normal samples of equal sizes with sample sizes starting at 10 per group and incremented by five per group at each iteration. Iterations terminated and the sample size recorded when the specified power (.50, .67 or .80) was attained or, if not attained, when a sample size of 1000 per group was reached.

For AIC and BIC, the proportion of cases for which the selection procedure resulted in selection of the correct data-generating model represents the true-model (or, accuracy) rate. For HSD, pairwise q tests were calculated for all pairs of means and a count was made of the number of correct decision in the sense of identifying the correct pattern (e.g., to be counted as correct for the population pattern {1, 2, 3, 4, 5}, all 10 pairwise differences had to be significant at the .05 level). Note that the simulations only involved equal sample sizes with equal population variances.

Results

Results for minimum sample sizes are shown in Tables 1, 2 and 3 for effect sizes of .10, .25 and .40, respectively. As expected from prior power studies, HSD often requires considerably larger sample sizes to attain specified power levels than do methods based on information criteria. However, there are substantive differences among the methods for specific cases. The following generalities apply:

(A) When all means are different, AIC requires uniformly much smaller sample sizes than either BIC or Tukey HSD for any number of groups. For example, this superiority of AIC is displayed in Figure 1 that shows minimum sample size requirements for AIC, BIC and Tukey HSD with medium effect size, .25, medium power, .67, and all means different. On the other hand, the minimum sample size requirements for BIC and Tukey HSD are essentially equivalent for this case.

(B) As a rule of thumb, AIC requires smaller minimum samples sizes than BIC or Tukey HSD when the number of clusters of homogeneous means is greater than one-half the number of groups. Occasionally this rule fails since AIC cannot, in theory, attain .67 or .8 power, as noted above.

(C) When the number of clusters of homogeneous means is less than one-half the number of groups, BIC tends to perform better than either AIC or Tukey HSD although this advantage tends to vanish when all group means are equal. On the basis of the poor performance of AIC for the null pattern, it was suggested by Dayton(1998) that an omnibus test be conducted as the first step in any analysis and that additional analyses be contingent on attaining significance with the omnibus test. However, a preliminary omnibus test provides no benefit for the BIC strategy.

(D) For three or more clusters of homogeneous means, those patterns with two or more groups clustered in the center yield higher accuracy rates than when the groups are clustered in the tail for all three methods. For

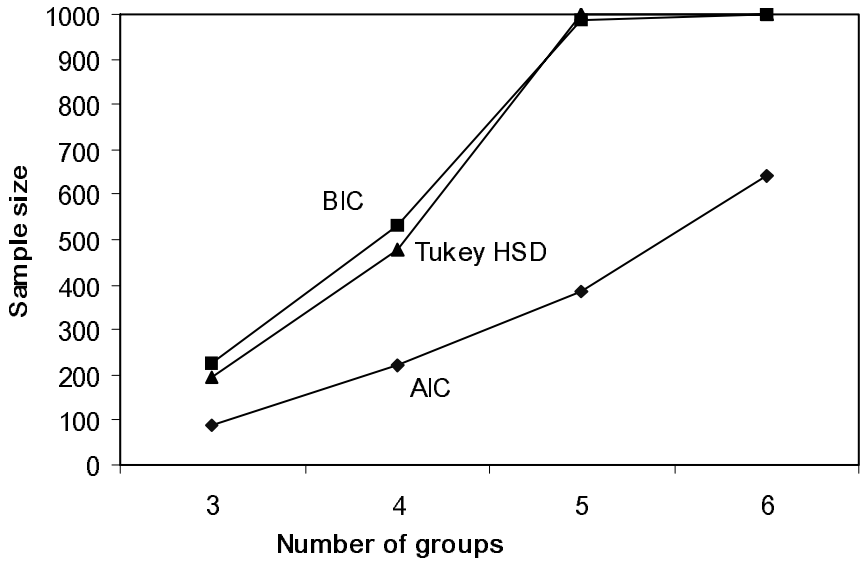

Figure1 All means different Power=. 67

example, with four groups, the pattern {1, 23, 4} has higher accuracy rates than pattern {12, 3, 4} even though both patterns contain three clusters of means. Similarly, for six groups, the five-cluster pattern {1, 2, 34, 5, 6} requires smaller minimum sample size requirements than the five-cluster pattern {12, 3, 4, 5, 6}.

In general, inconsistent performance between the two PCIC methods, AIC and BIC, can be explained by differences in their penalty terms. In general, AIC tends to select more complex models than BIC. Thus, when errors are made, AIC can be viewed as tending to overfit models whereas BIC can be viewed as tending to underfit models.

Table 1. Minimum Sample Size Requirement: Effect Size=0.10

	AIC			BIC			Tukey HSD		
Power	.5	.67	.8	.5	.67	.8	.5	.67	.8
Pattern of means									
Three groups									
{1,2,3}	560	750	985	M	M	M	M	M	M
{12,3}	100	185	325	225	310	415	345	450	565
{123}	10	10	*	10	10	10	10	10	10
Four groups									
{1,2,3,4}	M	M	M	M	M	M	M	M	M
{12,3,4}	615	860	M	M	M	M	M	M	M
{1,23,4}	390	550	835	910	M	M	M	M	M
{123,4}	110	220	*	175	245	325	370	480	580
{1234}	10	*	*	10	10	10	10	10	10
Five groups									
{1,2,3,4,5}	M	M	M	M	M	M	M	M	M
{12,3,4,5}	M	M	M	M	M	M	M	M	M
{12,3,45}	655	M	*	M	M	M	M	M	M
{1,234,5}	360	595	*	665	805	980	M	M	M
{1234,5}	105	*	*	135	205	260	385	495	575
{12345}	10	*	*	10	10	10	10	10	10
Six groups									
{1,2,3,4,5,6}	M	M	M	M	M	M	M	M	M
{12,3,4,5,6}	M	M	M	M	M	M	M	M	M
{1,2,34,5,6}	M	M	M	M	M	M	M	M	M
{1,2,3,456}	M	M	*	M	M	M	M	M	M
{1,2,345,6}	M	M	*	M	M	M	M	M	M
{12,34,56}	515	*	*	710	930	M	M	M	M
{12,345,6}	405	*	*	465	580	740	M	M	M
{12345,6}	160	*	*	125	170	230	385	465	545
{123456}	*	*	*	10	10	10	10	10	10

* AIC, cannot, in theory attain this power
M Sample size >1000

Table 2. Minimum Sample Size Requirement: Effect Size=0.25

	AIC			BIC			Tukey HSD		
Power	.5	.67	.8	.5	.67	.8	.5	.67	.8
Pattern of means									
Three groups									
{1,2,3}	90	125	160	225	265	325	195	240	285
{12,3}	20	30	60	30	45	60	60	75	90
{123}	10	10	*	10	10	10	10	10	10
Four groups									
{1,2,3,4}	220	275	335	530	640	730	480	575	655
{12,3,4}	100	145	235	210	255	310	250	305	360
{1,23,4}	60	90	125	120	155	185	200	230	280
{123,4}	20	45	*	25	35	50	65	80	95
{1234}	10	*	*	10	10	10	10	10	10
Five groups									
{1,2,3,4,5}	385	475	565	985	M	M	M	M	M
{12,3,4,5}	240	320	485	520	620	740	585	670	765
{12,3,45}	100	185	*	145	200	245	335	395	450
{1,234,5}	55	90	*	85	100	130	175	210	240
{1234,5}	20	*	*	20	25	40	60	75	90
{12345}	10	*	*	10	10	10	10	10	10
Six groups									
{1,2,3,4,5,6}	640	760	925	M	M	M	M	M	M
{12,3,4,5,6}	460	610	880	M	M	M	M	M	M
{1,2,34,5,6}	310	420	550	670	765	900	885	M	M
{1,2,3,456}	260	420	*	545	650	740	680	765	905
{1,2,345,6}	170	270	*	300	360	430	470	540	625
{12,34,56}	85	250	*	105	140	175	360	415	480
{12,345,6}	65	*	*	65	90	110	260	305	340
{12345,6}	30	*	*	20	25	40	65	75	90
{123456}	*	*	*	10	10	10	10	10	10

* AIC, cannot, in theory attain this power
M Sample size >1000

Table 3. Minimum Sample Size Requirement: Effect Size=0.40

	AIC			BIC			Tukey HSD	
Power	.5	.67	.8	.5	.67	.8	.5	.67	.8
Pattern of means									
Three groups									
{1,2,3}	40	50	60	75	95	115	80	90	120
{12,3}	10	15	25	10	15	25	25	30	40
{123}	10	10	*	10	10	10	10	10	10
Four groups									
{1,2,3,4}	85	105	135	195	230	275	190	220	260
{12,3,4}	40	55	85	75	90	110	105	125	145
{1,23,4}	25	40	60	45	55	70	80	90	105
{123,4}	10	15	*	10	15	20	25	30	35
{1234}	10	*	*	10	10	10	10	10	10
Five groups									
{1,2,3,4,5}	160	195	235	365	425	475	365	415	480
{12,3,4,5}	100	130	190	200	235	290	245	285	315
{12,3,45}	50	70	*	60	80	100	140	165	185
{1,234,5}	25	35	*	30	40	50	75	90	105
{1234,5}	10	*	*	10	15	15	25	35	40
{12345}	10	*	*	10	10	10	10	10	10
Six groups									
{1,2,3,4,5,6}	250	300	365	580	690	765	600	675	760
{12,3,4,5,6}	175	235	350	385	455	525	455	515	580
{1,2,34,5,6}	120	155	240	235	280	330	355	400	450
{1,2,3,456}	105	155	*	190	235	275	260	305	350
{1,2,345,6}	65	105	*	105	130	160	190	220	245
{12,34,56}	40	85	*	40	50	65	145	165	190
{12,345,6}	30	*	*	25	35	45	105	115	130
{12345,6}	15	*	*	10	10	20	25	30	40
{123456}	*	*	*	10	10	10	10	10	10

* AIC, cannot, in theory attain this power
M Sample size >1000

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B.N. Petrov & F. Csake (eds.), *Second International Symposium on Information Theory*. Budapest: Akademiai Kiado, 267-281.

Aptech Systems, Inc. (2002). *Gauss for Windows version 4*. Maple Valley, WA

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*, 345-370.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Cribbie, R. A. & Keselman, H. J. (2003). Pairwise multiple comparisons: A model comparison approach versus stepwise procedures. *British Journal of Mathematical & Statistical Psychology*, *56*, 167-182.

Cribbie, R. A. (2003). Pairwise multiple comparisons: New yardstick, new results. *The Journal of Experimental Education*, *71*, 251-265.

Dayton, C. M. (1998). Information criteria for the paired-comparisons problem. *The American Statistician*, *52*, 144-151.

Dayton, C. M. (2003). Information criteria for pairwise comparisons. *Psychological Methods*, *8*, 61-71.

Ramsey, P. H. (1978). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, *52*, 333-343.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461-464.